IN THE SPECIFICATION

Please replace the Title of the Application with the following amended Title:

Method and System for Generating Training Data for an Automatic Speech ~~Recogniser~~ Recognizer

Please replace the paragraph at page 1, lines 6-10 with the following amended paragraph:

This invention relates in general to a method and system for generating training data for an automatic speech recognizer ~~recogniser~~ operating at a particular sampling frequency. Furthermore, the invention relates to a method for training an automatic speech recognition system, and a method and system for generating a codebook for use in the method for generating the training data.

Please replace the paragraph at page 1, lines 11-26 with the following amended paragraph:

Automatic speech recognizers ~~recognisers~~ are used for various applications such as control interfaces, automobile navigation systems, dialog systems etc., in which speech input is identified and interpreted. Generally, the user of such an automatic speech recognizer ~~recogniser~~ (ASR) speaks into a microphone, where the analog speech input is converted into digital form by the usual techniques of windowing and sampling the input analog signal, i.e. measuring the amplitude of the analog signal at a continuous rate to give a set of discrete samples. The rate at which the signal is sampled is called the sampling rate or sampling frequency. The resulting sequence of discrete samples gives a time-domain description of the analog signal. This time-domain description of the input signal is converted to a frequency domain description, for example by performing a Fast Fourier Transform on the sampled input signal, where various processing steps are performed to extract features, often in the form of feature vectors, for the input signal. By comparing these features to templates or other models, referred to in the following as "models", and locating a most suitable match, the ASR is able to analyze ~~analyse~~ the speech input to determine what the user has said and which actions are to be carried out as a result.

Please replace the paragraph at page 1, line 27 – page 2, line 7 with the following amended paragraph:

The models used by an automatic speech recognizer ~~recogniser~~ are usually computed using training data, which are generally a collection of spoken utterances such as words, sentences or entire conversations. The training data are input to a front end, i.e. the first processing stages, of an automatic speech recognizer ~~recogniser~~ and processed to calculate the models for the automatic speech recognizer ~~recogniser~~. To increase the success rate of the automatic speech recognizer ~~recogniser~~ in correctly identifying and understanding the input speech during operation, it is usual to employ a number of speakers for the training of the automatic speech recognizer ~~recogniser~~ with differing accents or intonations to give as broad a selection of utterances as possible. The more utterances available for training the of the automatic speech recognizer ~~recogniser~~, the better its performance. Even better performance is attained if the training data are recorded under acoustic conditions similar to the conditions in which the automatic speech recognizer ~~recogniser~~ is intended to operate.

Please replace the paragraph at page 2, lines 8-20 with the following amended paragraph:

Every analog signal can be regarded as a composite of many component sinusoidal waves of different frequencies. The sampling frequency is chosen according to the desired quality of the samples. A high sampling rate ensures that higher-frequency components are included in the sampled signal. According to Nyquist, the sampling frequency must be at least twice the frequency of the highest desired frequency component, since any component frequency lower than half of the sampling rate is lost in sampling. Therefore, an automatic speech recognizer ~~recogniser~~ will benefit from a higher sampling rate for the input speech, due to additional information in the higher frequency bands which improves recognition of the speech. For example, an automatic speech recognizer ~~recogniser~~ operating in an automobile can perform considerably better at a higher sampling rate. To train a such an automatic speech recognizer ~~recogniser~~ operating at a higher sampling rate, it is necessary to first collect training audio data acquired at this sampling rate.

Please replace the paragraph at page 2, line 21 – page 3, line 5 with the following amended paragraph:

Training data for an automatic speech recognizer ~~recogniser~~ should cover as wide a variety of spoken utterances as possible, for example single words, whole sentences, or even entire conversations. Ideally, the spoken words originate from a number of speakers with different accents or quality of articulation. Therefore, to collect sufficient raw data to calculate the necessary number of models for robust performance of the automatic speech recognizer ~~recogniser~~, it would require many persons recording large numbers of test words under realistic conditions to reflect the noisy environment of a typical automobile. With an appropriately diverse collection of utterances, the training data can ensure robust operation of the automatic speech recognizer ~~recogniser~~, with reliable recognition of speech under actual working conditions. However, training audio data for automatic speech recognizers ~~recognisers~~ operating at higher sampling frequencies, e.g. for an automatic speech recognizer ~~recogniser~~ for use in automotive applications, are not readily available, since collecting data in adverse environments such as in a noisy automobile is very time-consuming and therefore prohibitively expensive. Furthermore, each type of automatic speech recognizer ~~recogniser~~ requires training data in the form of feature models in its own specific format. Training data in a format for a particular make of ASR may be quite unsuitable for a different type of ASR.

Please replace the paragraph at page 3, lines 6-8 with the following amended paragraph:

Therefore, an object of the present invention is to provide an easy and inexpensive method and system for generating training data for any automatic speech recognizer ~~recogniser~~.

Please replace the paragraph at page 3, lines 9-14 with the following amended paragraph:

To this end, the present invention provides a method for generating training data for an automatic speech recognizer ~~recogniser~~ - constructed for a particular first sampling frequency--by deriving spectral characteristics from audio data sampled at a second frequency lower than the first sampling frequency, extending the bandwidth of the spectral characteristics by retrieving bandwidth extending information, and processing the bandwidth extended spectral characteristics to give the required training data.

Please replace the paragraph at page 3, lines 20-30 with the following amended paragraph:

An inherent advantage of this method is that the training audio data used to generate the training data might be data that are already available for use in other, different applications, and might have been sampled at a lower frequency than required for the training data. Therefore, for example, databases of available telephone audio data might be implemented, since such databases are already available, are generally quite large, and cover a wide variety of spoken words and/or entire sentences from, typically, diverse sets of speakers. Since the bandwidth of 4 kHz generally suffices for telephony use, audio telephone data is usually sampled at 8 kHz. With the method according to this invention, this 8 kHz data may be used for the training of an automotive automatic speech recognizer ~~recogniser~~ which, for reasons of performance quality, might operate at a relatively higher frequency such as 11 kHz or even higher.

Please replace the paragraph at page 3, line 31 – page 4, line 5 with the following amended paragraph:

An appropriate system for generating training data for an automatic speech recognizer ~~recogniser~~ operating at a particular first sampling frequency comprises a converter for deriving spectral characteristics from audio data sampled at a second frequency lower than the first sampling frequency , a retrieval unit for retrieving bandwidth extending information for the spectral characteristics from a codebook, and a processing module for processing the bandwidth-extended spectral characteristics to give the required training data.

Please replace the paragraph at page 4, lines 6-18 with the following amended paragraph:

According to the present invention, the bandwidth of the spectral characteristics of the data available at the lower sampling frequency is extended so that the input appears to have been sampled at the higher frequency. The bandwidth extending information can be retrieved from a suitable source where it is stored, in an appropriate form. Here, such a source is commonly called a "codebook". A codebook is therefore a collection of templates or stochastic mixture models in a certain form, to which other data, in the same form, can be compared. The form of the data is generally quite complex, for example, the feature vectors for a typical ASR may be n-dimensional vectors, where n is often quite a large number, and comparison of the data to the templates usually involves locating the "best fit". This codebook, used to generate training data for an automatic speech recognizer ~~recogniser~~, is not to be confused with a different type of codebook which may be used in later stages of the automatic speech recognizer ~~recogniser~~, and which is of no relevance here.

Please replace the paragraph at page 4, lines 19-21 with the following amended paragraph:

The bandwidth extended spectral characteristics can then be processed in a next step to give the training data in a form required by further stages of the automatic speech recognizer ~~recogniser~~.

Please replace the paragraph at page 6, lines 19-26 with the following amended paragraph:

Spectral characteristics might also be obtained by calculating the logarithm of the filterbank power values in a further processing step to give a set of log-spectral coefficients. In the case where a warping of the frequency axis is effected in the filterbank according the mel scale, the resulting coefficients can be referred to as mel frequency coefficients. Such log-spectral coefficients are often the basis for generating feature vectors for use in systems such as automatic speech recognizers ~~recognisers~~. The log-spectral coefficients might also be calculated using a different, equally suitable, technique.

Please replace the paragraph at page 6, line 27 – page 7, line 1 with the following amended paragraph:

In a particularly preferred embodiment of the invention, the log-spectral coefficients are used as the spectral characteristics for generating entries for the bandwidth extension codebook for use in a system for generating training data for an automatic speech recognizer ~~recogniser~~. By performing a DCT on the log-spectral coefficients, these can be transformed into log-cepstral coefficients, which are particularly suited for application in the further processing steps of the automatic speech recognizer ~~recogniser~~ such as speech identification and understanding.

Please replace the paragraph at page 7, lines 2-7 with the following amended paragraph:

In an appropriate method for training an automatic speech recognizer ~~recogniser~~, it is sufficient to generate the required training data using audio data sampled at a lower frequency and augmented with bandwidth extending information retrieved from a codebook, giving training data which appear to have been obtained at a higher sampling frequency. Nevertheless, the training data sampled at a lower frequency could be used along with training data sampled at the required frequency.

Please replace the paragraph at page 7, line 30 – page 8, line 6 with the following amended paragraph:

Similarly, the audio data sampled at the lower frequency and used to generate training data for the automatic speech recognizer ~~recogniser~~ may also require spectral modification to remove unwanted noise or channel effects. Such spectral features present in the audio data might have a negative effect when incorporated into the training data, and are preferably removed by continuously calculating the average or mean spectrum from the audio data and subtracting the mean spectrum from the spectral characteristics of the audio data before retrieving the bandwidth extending information from the codebook. This ensures that the training data generated for the automatic speech recognizer ~~recogniser~~ is essentially free of unwanted noise or channel effects.

Please replace the paragraph at page 8, lines 7-17 with the following amended paragraph:

Since the training data for the automatic speech recognizer ~~recogniser~~ should realistically reflect the typical audio qualities of the environment in which it is intended to operate, it may be desirable to add in or insert suitable background noise information, or other similar spectral features. To this end, the spectrum of the bandwidth extended spectral characteristics might be adjusted to alter its spectral properties in an optional processing step. For reasons of computational ease, such a processing step is preferably carried out in the linear domain. This might necessitate a step of calculating the inverse log of the spectral characteristics, should these be in logarithmic form. The spectrum of the audio data can then be modified by adding in the required features. The logarithm of the spectrum is then calculated again, as necessary, to return the spectrum to the log domain.

Please replace the paragraph at page 8, lines 22-23 with the following amended paragraph:

FIG. 1 is a block diagram showing usual processing steps in a front end of an automatic speech recognizer ~~recogniser;~~

Please replace the paragraph at page 8, lines 26-27 with the following amended paragraph:

FIG. 3 is a block diagram of a system for generating training data for an automatic speech recognizer ~~recogniser~~ according to an embodiment of the invention;

Please replace the paragraph at page 8, line 32 – page 9, line 16 with the following amended paragraph:

In FIG. 1, a simplified representation shows the stages in a typical front end of an automatic speech recognizer ~~recogniser~~ involved in processing an input analog audio signal A to generate feature vectors V for the audio signal for use at a later stage in speech recognition. The analog audio signal A, which may comprise both speech and noise components, is first windowed and sampled at a sampling frequency f to give sets of digital audio samples. A Fast Fourier Transform (FFT) is performed for each set of digital samples, giving a corresponding set of Fourier coefficients. These in turn are forwarded to a filterbank in which the filters are configured in a non-linear manner according to the Bark or mel scale, to calculate the energies of the signal's various frequency components, giving a set of filterbank energy values. The logarithm is calculated for the filterbank energy values in a log unit to give a set of log filterbank coefficients. A Long Term Normalisation (LTN) is performed on the log filterbank coefficients in order to normalise channel effects. The LTN output is then further processed by performing a Discrete Cosine Transform (DCT) on the log spectrum coefficients to give feature vectors V, in this case cepstral coefficients. In further stages of the automatic speech recognizer ~~recogniser~~, not shown in this diagram, the feature vectors V are used for speech recognition and speech understanding.

Please replace the paragraph at page 9, lines 17-23 with the following amended paragraph:

FIG. 2 shows a system for generating a codebook 6 for use in a system according to FIG. 3 for generating training data for an automatic speech recognizer ~~recogniser~~ 2 built for a sampling frequency f_H, and which is to be trained using data sampled at a lower frequency f_L. Audio data DC_H, which has already been sampled at the higher frequency f_H, are processed by a module 9, similar in parts to the front end of an automatic speech recognizer ~~recogniser~~ as described in FIG. 1. At the same time, the audio data are processed by a similar module 10. Modules 9 and 10 can be seen in detail in FIGS. 4 and 5 respectively.

Please replace the paragraph at page 10, line 27 – page 11, line 3 with the following amended paragraph:

How this codebook could be used to generate training data for an automatic speech recognizer ~~recogniser~~ $2(f_H)$ which is built to operate on a sampling frequency f_H can be seen in FIG. 3. The input audio data D_L for training the automatic speech recognizer ~~recogniser~~ $2(f_H)$ is available at a lower frequency f_L. The input audio data D_L are first processed in a module 3 similar in parts to the front end of an automatic speech recognizer ~~recogniser~~ as already described in FIG. 1 to give sets of spectral characteristics S_L. The module 3 is constructed in a manner identical to module 9, used to process the audio data in the codebook generation process described in FIG. 2. This shows that the audio data at higher sampling frequency are processed in the same manner in both cases.

Please replace the paragraph at page 12, lines 25-32 with the following amended paragraph:

In order to reflect the environment in which the automatic speech recognizer ~~recogniser~~ 2 will operate, the spectral characteristics $S_{L,E}$ can be modified accordingly in an optional block 8. This optional block 8 is shown here as part of the final processing module 7, placed before the DCT. For example, noise can be added to the spectrum to reflect the noisy environment in an automobile. Since this type of operation should be performed in the linear spectral domain, the inverse log is first calculated for the spectral characteristics $S_{L,E}$ before adding the noise spectrum and calculating the logarithm for the spectral characteristics $S_{L,E}$ once again.